Journal of Computational Mathematics and Data Science 13 (2024) 100103

Contents lists available at ScienceDirect

Journal of Computational Mathematics and Data Science

journal homepage: www.elsevier.com/locate/jcmds

Enhanced MRI brain tumor detection and classification via topological data analysis and low-rank tensor decomposition

Serena Grazia De Benedictis [a,1], Grazia Gargano [a,b,1,*], Gaetano Settembre [a,1]

[a] Department of Mathematics, University of Bari Aldo Moro, Bari, Italy
[b] Hematology and Cell Therapy Unit, IRCCS Istituto Tumori "Giovanni Paolo II", Bari, Italy

ARTICLE INFO

Keywords:
Brain tumor classification
Magnetic resonance imaging
Topological data analysis
Machine learning
Tucker decomposition

ABSTRACT

The advent of artificial intelligence in medical imaging has paved the way for significant advancements in the diagnosis of brain tumors. This study presents a novel ensemble approach that uses magnetic resonance imaging (MRI) to identify and categorize common brain cancers, such as pituitary, meningioma, and glioma. The proposed workflow is composed of a two-fold approach: firstly, it employs non-trivial image enhancement techniques in data preprocessing, low-rank Tucker decomposition for dimensionality reduction, and machine learning (ML) classifiers to detect and predict the type of brain tumor. Secondly, persistent homology (PH), a topological data analysis (TDA) technique, is exploited to extract potential critical areas in MRI scans. When paired with the ML classifier output, this additional information can help domain experts to identify areas of interest that might contain tumor signatures, improving the interpretability of ML predictions. When compared to automated diagnoses, this transparency adds another level of confidence and is essential for clinical acceptance. The performance of the system was quantitatively evaluated on a well-known MRI dataset, with an overall classification accuracy of 97.28% using an extremely randomized trees model. The promising results show that the integration of TDA, ML, and low-rank approximation methods is a successful approach for brain tumor identification and categorization, providing a solid foundation for further study and clinical application.

1. Introduction

Medical imaging is the technique for visual representing the structure and function of parts of the human body for clinical analysis and medical purposes. Its primary objective is to study the normal anatomy and physiology of different tissues and organs while also enabling the identification of abnormalities [1]. The field of healthcare science has been revolutionized by medical imaging. Registration of multimodality images is particularly important for diagnosing of abnormalities associated with tumors, planning surgical and radiological treatment, tracking changes in tissue morphology associated with disease progression or response to therapy, and relating anatomical information to changes in functional characteristics [2]. Magnetic resonance imaging (MRI) is primarily an imaging technique used to non-invasively visualize the anatomy and physiology of the body in both disease and health. MRI also provides excellent contrast of soft body tissues; for example, the white and gray matter structure of the brain can be easily distinguished. For this reason, MRI is often used to analyze brain tissue according to size, shape or location, and this helps to detect tumors [3]. A brain tumor occurs when abnormal cells form in the brain, which can be either malignant (cancerous) or benign.

* Corresponding author at: Department of Mathematics, University of Bari Aldo Moro, Bari, Italy.
E-mail addresses: serena.debenedictis@uniba.it (S.G. De Benedictis), grazia.gargano@uniba.it (G. Gargano), gaetano.settembre@uniba.it (G. Settembre).
[1] All authors contribute equally to this work.

https://doi.org/10.1016/j.jcmds.2024.100103
Received 27 June 2024; Received in revised form 3 September 2024; Accepted 29 September 2024

Table 1
Distribution of samples in training and test sets.

Class	Number of images	Training set	Testing set
	7023	5712	1311
Glioma	1621	1321	300
Meningioma	1656	1339	306
Pituitary	1757	1457	300
No tumor	2000	1595	405

Primary brain tumors start in the brain itself or in surrounding tissue. Some of the most common primary tumors seen in clinical practice are meningiomas, pituitary tumors and gliomas. Meningiomas arise from the meninges and compress the brain as they grow. Pituitary tumors originate from the cells of the pituitary gland, a vital gland located at the base of the brain near the optic chiasm. Brain gliomas, including the most common type, glioblastoma, originate from glial cells within the brain parenchyma. The incidence of primary brain tumors varies with age, gender and ethnicity. Malignant brain tumors, such as gliomas, are slightly more common in men, while meningiomas and pituitary are generally more common in women [4].

Early detection of brain tumors plays an important role in improving the effectiveness of treatment and further increasing patient survival rates. However, the study of brain tumors using medical imaging is sometimes complex. In addition, manual segmentation of brain tumors is costly and time-consuming. Automated approaches are therefore highly valued. Automated detection of brain tumors is a challenging medical task. In recent years, many techniques ranging from simple machine learning (ML) models [5–7] to more recent state-of-the-art techniques such as convolutional neural networks (CNN) [8–10] and vision transformers (ViT) [11,12] have been used to automatically classify brain malignancies using brain MRI [13]. However, the implementation of a translational and explainable model for the accurate detection and categorization of brain tumors remains an active research task.

In this article, we propose a novel framework for brain tumors classification that combines dimensionality reduction approach with ML algorithms, and then merges ML prediction with topological data analysis (TDA)-based methods. After image preprocessing steps which are discussed in detail in the following sections, low-rank Tucker decomposition (TD) [14] effectively reduces the data dimensionality while preserving essential features and structures. This reduction not only reduces the computational burden, but also facilitates feature extraction, potentially revealing intricate patterns and relationships within the data that may be critical for tumor classification. Subsequently, the use of supervised ML classifiers trained on a low-dimensional representation of the data enables the development of robust tumor classification models. We perform multi-class classification of MRI brain scans using different ML classifiers. Unlike conventional approaches, our approach integrates normal MRI scans (i.e., brain images that do not contain tumors) into the classification task, with the aim of increasing the precision and robustness of the categorization process. By including normal MRI scans, our approach provides a more comprehensive and holistic model for identifying brain abnormalities. The medical image data is processed in parallel using a unique approach that uses TDA through persistent homology (PH) to identify regions of interest (ROI) in the MRI scans. This comprehensive approach improves the interpretability of the classification process and could also potentially provide deeper insights into tumor characteristics. The aim is to improve the diagnosis and treatment of brain tumors, potentially revolutionizing future screening methods to make them more efficient and effective.

The remainder of the paper is structured as follows. Section 2 provides an overview of the materials and methods used in our research. In Section 3 we present our proposed approach and describe its implementation. Section 4 discusses the results obtained, and finally, Section 5 concludes the article with some final remarks.

2. Materials and methods

2.1. Dataset

The MRI dataset[2] contains 7023 human brain MRI scans of patients mainly classified into four different categories, namely glioma, meningioma, pituitary, and no tumor. Out of these images, 5712 images are put into the training set and the remaining 1311 images are put into the test set, according to the guidelines of the Kaggle challenge. The dataset contains a reasonably balanced number of samples per class: the balanced distribution between classes prevents the models from favoring any predominant class, thus ensuring unbiased and accurate performance across all categories. Detailed information on the number of samples per class are reported in the Table 1.

The MRI dataset is a combination of the following three datasets: SARTAJ [15], figshare [16], and Br35H [17]. MRI head scans were acquired in different orientations: axial, coronal, sagittal, as shown in Fig. 1. It should be observed that the MRI scans lack labels indicating their orientation, and the dimensions of the images can vary significantly. These intrinsic characteristics of the data pose substantial challenges for both preprocessing and ML modeling.

In addition, the morphological characteristics of each tumor type are very different [18]. In particular, meningiomas generally present well-defined, rounded masses, often appearing more uniform in texture compared to other tumor classes. The glioma tumor

[2] The dataset was collected online and it is publicly available at the following link: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset.

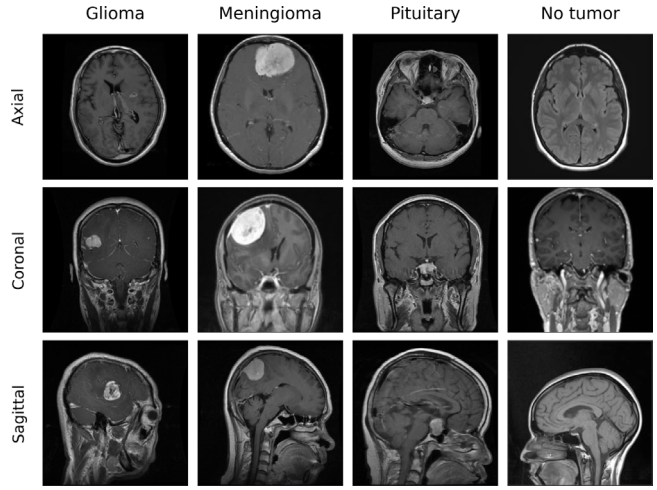


Fig. 1. Examples of MRI scans in the dataset. For each class, the samples acquired in different planes, are reported.

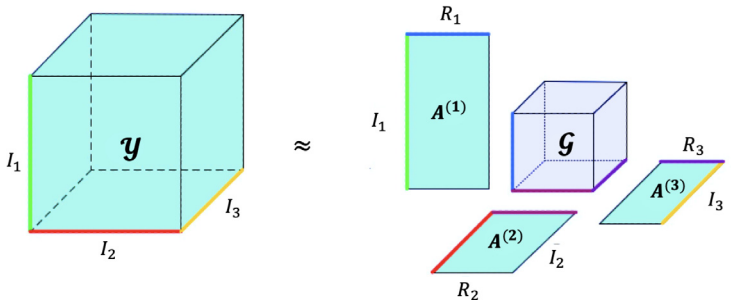


Fig. 2. Low-rank TD for a 3-dimensional tensor.

typically presents irregularly shaped masses with heterogeneous intensity, making it one of the more challenging tumors to identify. Images labeled as pituitary tumors usually show smaller, more localized masses near the base of the brain, which often appear to be distinct from the surrounding tissue.

2.2. Methods

In this section, we provide a comprehensive theoretical overview of the methods and techniques used in the study. In particular, we discuss the building blocks of the proposed approach, including the low-rank TD, the ML algorithms employed, and the TDA techniques.

2.2.1. Low-rank tucker decomposition

Treating data in tensor form allows the use of tensor decomposition to achieve size reduction while preserving the critical information contained within the data. In fact, this reduction not only alleviates computational burdens, but also facilitates extraction of features and patterns within the data that may be crucial for tumor classification. Among the various existing tensor decompositions, we use the TD that reveals the most suitable for its effectiveness in ML modeling and its ability to provide a more compact and interpretable representation of multidimensional data. TD [14] factorizes a K-order tensor $\mathcal{Y} \in \mathbb{R}^{I_1 \times \cdots \times I_K}$ into a core tensor $\mathcal{G} \in \mathbb{R}^{R_1 \times \cdots \times R_K}$ and K factor matrices $A^{(i)} \in \mathbb{R}^{I_i \times R_i}$. Without loss of generality, the factor matrices can be assumed to have orthonormal columns. The input tensor can therefore be approximated with:

$$\mathcal{Y} \approx \mathcal{G} \times_1 A^{(1)} \times_2 \ldots \times_K A^{(K)}, \tag{1}$$

where $\times_n$ denotes the mode-n matrix product, which essentially projects data onto given basis factors. Fig. 2 illustrates the TD of a third-order tensor graphically. If $R_i < I_i$, the core tensor $\mathcal{G}$ can be seen as a compressed version of the original tensor. Instead, the factor matrices $A^{(i)}$ play a crucial role in this decomposition: their column vectors represent the set of basis functions onto which the data is projected and define the mapping between the original data and its compressed representation (core tensor), and vice versa.

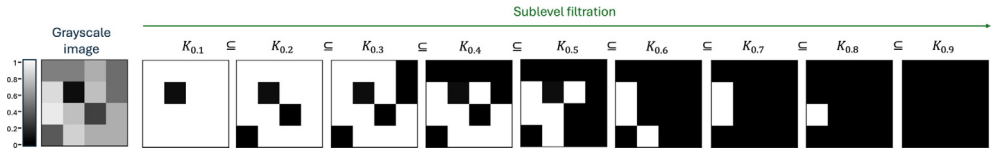


Fig. 3. Graphical representation of sublevel filtration on 4 × 4 subregion of an MRI scan.

The TD problem of decomposing a data tensor $\mathcal{Y} \in \mathbb{R}^{I_1 \times \cdots \times I_K}$ can be formulated as

$$\min_{\mathcal{G}, A^{(1)}, \ldots, A^{(K)}} \|\mathcal{Y} - \mathcal{G} \times_1 A^{(1)} \times_2 \cdots \times_K A^{(K)}\|, \tag{2}$$

$$\text{subject to } \mathcal{G} \in \mathbb{R}^{R_1 \times \cdots \times R_K},$$
$$A^{(n)} \in \mathbb{R}^{I_n \times R_n} \quad \text{and column-wise orthogonal for } n = 1, \ldots, K.$$

To compute the TD, we utilize the higher-order orthogonal iteration (HOOI) algorithm [19].

2.2.2. Machine learning approaches

Given the labeled nature of the MRI dataset, the corresponding classification problem can be tackle with different well-known ML algorithms able to produce six distinct classifiers to accurately categorize brain tumor images. Specifically, we chose classical ML methods over deep learning approaches due to their improved interpretability and efficiency, which are particularly suited to smaller datasets. This decision highlights the continued effectiveness of classical ML techniques in meeting specific domain requirements, providing robust performance without excessive computational demands. The ML classifiers selected include k-nearest neighbor (KNN) [20], support vector machine (SVM) [21], random forest (RF) [22], extremely randomized trees (also known as extra-trees) [23], extreme gradient boosting (XGBoost) [24] and adaptive boosting (AdaBoost) [25]. To determine each classifier's effectiveness in the context of brain tumor detection and classification, a thorough examination and comparison will be conducted. In Section 4, the assessment metrics are described in detail.

2.2.3. Topological data analysis

ROI selection is a common step in medical image analysis across all imaging modalities. This area represents a subset of an image that is suitable for the intended analysis and is usually identified manually by experts. Manually analyzing such a large amount of data is challenging and time-consuming, but essential for accurate analysis. Automatic detection of regions of interest can significantly reduce the size of the image to be processed, speed up analysis, improve accuracy and assist domain experts.

Inspired by earlier findings in the literature [26], we propose an approach that uses TDA to enable the interpretable and unsupervised identification of a relevant region in an MRI brain scan. TDA is a field of data analysis that uses concepts and methods from algebraic topology to study the shape and inner structure of data [27]. This approach is particularly useful when working with complex and high-dimensional data, such as images [28].

Employing the TDA technique known as PH [29], we can monitor the evolution of topological invariants (e.g., connected components, loops, and other topological patterns) through a filtration process, with the ultimate goal of identifying the most persistent invariant that provides meaningful information about the analyzed data. This study illustrates the efficacy of PH to identify, in an unsupervised manner, persistent connected components within MRI scans. The innovative aspect of this approach resides in the extraction of topological features within the image – through a technique rooted in mathematical theory – which detect pathological area of interest and distinguishes it from the non pathological regions of the brain, where pixel intensities are generally uniform in color. Since this approach is rooted in fundamental concepts of pure mathematics, the strategy for analyzing a given set of data involves the construction of a geometric structure called a cubic complex. The cubic complex generated over the data allows the use of tools from algebraic topology for computational purposes.

Given an image, this geometric structure can be obtained by exploiting the natural image pixel configuration [30]: in particular, each pixel is treated as a point (0-simplex), and a line segment (1-simplex) is established between two points if their corresponding pixels are adjacent. The evolution of topological invariants is monitored using a sublevel filtration applied to the constructed cubic complex of a fixed image. This filtration generates a nested sequence of cubic complexes by thresholding the grayscale values of the image, from the smallest to the largest one. Fig. 3 depicts the overall idea and presents an example performed over a 4 × 4 subregion of an MRI scan.

To elucidate and explicate the filtration process employed and illustrated in Fig. 3, we denote by $\{K_i\}_i$ the sequence of cubic complexes obtained from the sublevel filtration, being i the chosen grayscale value. Starting with the grayscale value 0.1 (the darkest), $K_{0.1}$ consists of a single pixel, which forms our initial/first connected component. As the grayscale value increases to 0.2, new pixels appear in the cubic complex $K_{0.2}$, with only one adjacent to the pixel from $K_{0.1}$, thereby extending the previously identified connected component. The other pixel, isolated, forms a separate connected component. Increasing the gray value, more pixels are created, but the number of connected components remains two until the gray value 0.6 is reached. At $K_{0.6}$, the second connected component merges with the first formed before. This resulting connected component persists as the only connected

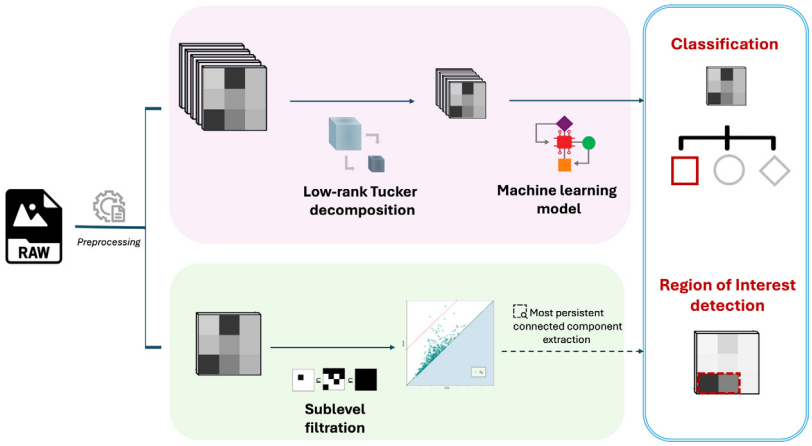


Fig. 4. Workflow of the proposed approach for brain tumor classification and ROI identification in MRI scans. After preprocessing, the data undergo parallel processing: one path involves dimensionality reduction and ML (pink block), while the other uses TDA (green block). The final output is the ML classifier's prediction and ROI identification.

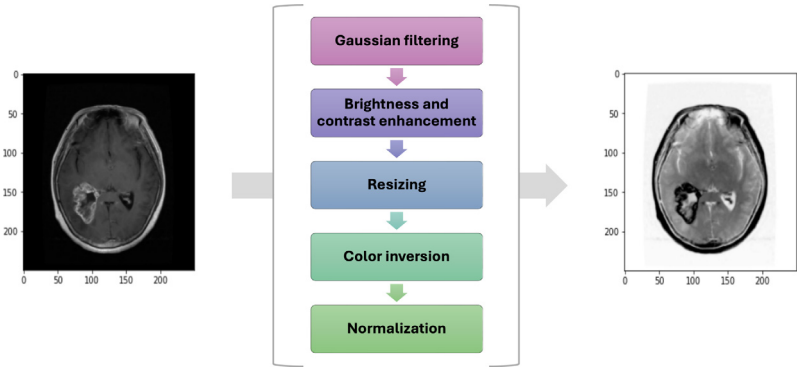


Fig. 5. Data flow diagram of preprocessing steps.

component throughout the rest of the filtering process. Therefore, the most persistent connected component characterizing the region of interest originates from $K_{0.1}$.

Information about the evolution of topological invariants in the analyzed data is captured in an algebraic structure known as a persistence diagram (PD) [31]. This is a concise representation of topological information that can be manipulated to develop a method for computationally and automatically extracting the most persistent connected component from the MRI scan, as discussed in the following.

3. Proposed approach

In this section, we describe both the proposed approach and the implementation details. First, we outline the data preparation process, focusing on the preprocessing of raw images; second, we provide insights into the computational aspects of our innovative methodology designed for diagnostic decision support of brain tumors. The entire proposed workflow, shown in Fig. 4, has been implemented using Python 3 programming language. Further details of additional libraries and code used to support the results of this study are available at https://github.com/gaetanosettembre/BrainTumor_clf_TDA.

3.1. Data preparation

Medical image preprocessing steps used in this study are depicted in Fig. 5. These steps are necessary for improving the image quality, eliminating noise, and emphasizing significant features that are critical for precise tumor identification and categorization. In particular, five preprocessing procedures are used as described in the following.

1. **Filtering**. MRI scans are susceptible to noise from magnetic radiation, mitigating this noise is of paramount importance. We applied two-dimensional Gaussian filter to noise removal and image smoothing while maintaining the integrity of the image

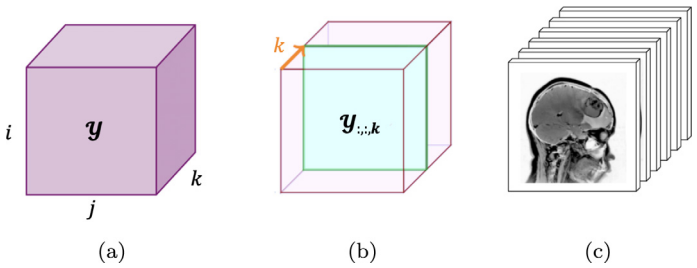


Fig. 6. Tensor data representation. (a) A 3-dimensional tensor. (b) Frontal slices of a 3-dimensional tensor. (c) Tensor obtained from MRI tumor scans.

edges [32]. This filter is defined as follows:

$$G(x,y) = \frac{1}{\sqrt{2\Pi\sigma^2}} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{3}$$

where σ, the standard deviation of the Gaussian distribution, controls the amount of smoothing. The Gaussian filter is a non-uniform linear filtering technique and applies a convolution operation with a Gaussian kernel to the input image. The kernel coefficients are symmetric in all directions and decrease in magnitude as the distance from the center of the kernel increases. Therefore, pixels in the center of the kernel carry higher weights compared to those at the edges. The kernel size of 5×5 pixels is used, while the σ value is automatically calculated according to the kernel size.

2. **Image Enhancement**. To improve the visual representation of the input image and convert it into a format appropriate for analysis, we modify the brightness and contrast characteristics. In particular, we used higher contrast and brightness values first to mitigate the potential reduction of image brightness caused by the Gaussian filter and second, to draw attention to the distinctions between the abnormal regions of the image – which may contain tumors marked by dyschromatic regions – and the non-pathological, typically uniformly colored brain regions.
3. **Resizing**. To structure heterogeneous image sizes into a cohesive tensor format, we re-scaled all images in the dataset to a size of 250×250. This procedure adjusts the width, the height or both, simultaneously.
4. **Color inversion.** This step involves changing each pixel's intensity so that the brightest regions turn into the darkest, and vice versa. Considering the grayscale image recorded in MRI scans for brain tumors, inversion modifies the image by flipping the grayscale values. We used color inversion to perform PH so that it may be easier to identify the most persistent connected component within the MRI.
5. **Normalization**. Pixel values are re-scaled in the standard range interval $[0, 1]$ to reduce the variability caused by different acquisition parameters for different images, improving the performance of ML models by guaranteeing that every feature contributes fairly to the analysis.

3.2. Implementation of the ensemble framework

The previous steps produce preprocessed MRI scans, now ready to be organized into tensors. The training and the test sets are two 3rd-order tensors, with the MRI scans stored as frontal slices, as shown in Fig. 6(c). In particular, the tensor $\mathcal{Y}$ containing the training images is of size $250 \times 250 \times 5712$, while the testing tensor $\overline{\mathcal{Y}}$ is of size $250 \times 250 \times 1311$.

To train ML algorithms with a low-rank dimensionality representation of data, we started by performing TD of the tensor $\mathcal{Y}$ of all sampling training data. From Eq. (1), it is clear that tensor decomposition perform data reduction by projecting the tensor $\mathcal{Y}$ to smaller dimension core tensors $\mathcal{G}$, where entries of the core tensor $\mathcal{G}$ are features of the training data $\mathcal{Y}$ in the feature space spanned by factors $A^{(n)}$ $(n = 1, 2, 3)$.

We chose the core dimensions R_1, R_2, R_3 experimentally by maximizing the performance metrics of the ML models. Further details regarding the selected best ranks values are provided in the subsequent section. We directly used for classification the training features obtained by projecting the core tensor $\mathcal{G}$ onto the basis factor $A^{(3)}$, i.e., $\mathcal{G}^{proj} = \mathcal{G} \times_3 (A^{(3)})^\top$. In this way, it is possible to add the class label information to the feature core tensor, which is necessary to training ML model. We then projected the test samples $\overline{\mathcal{Y}}$ using the basis factors $A^{(1)}$ and $A^{(2)}$ obtained from the training data. We denote by A^+ the Moore–Penrose inverse of a matrix A. The projected tensor $\overline{\mathcal{Y}}^{proj}$ is defined as $\overline{\mathcal{Y}}^{proj} = \overline{\mathcal{Y}} \times_1 (A^{(1)})^+ \times_2 (A^{(2)})^+$. Finally, we used the fitted ML classifiers to predict the classes of no tumor, meningioma, glioma, and pituitary for brain images of projected testing set. Fig. 7 illustrates the workflow for classification task.

The prediction generated by the ML classifier constitutes one of the outputs of our proposed approach. As depicted in Fig. 4, this output is complemented by the identification of a pertinent ROI in MRI scans. This non-trivial task is automatically accomplished through PH. Fig. 8 illustrates the steps performed to obtain ROI starting from pre-processed images. Particularly, the following operations are executed:

1. **Smoothing and cropping**. For each pixel in the image under investigation, we examined a square pixel neighborhood of size $N \times N$, and generate a new image by substituting the pixel with the average of its surrounding pixels. We employed a

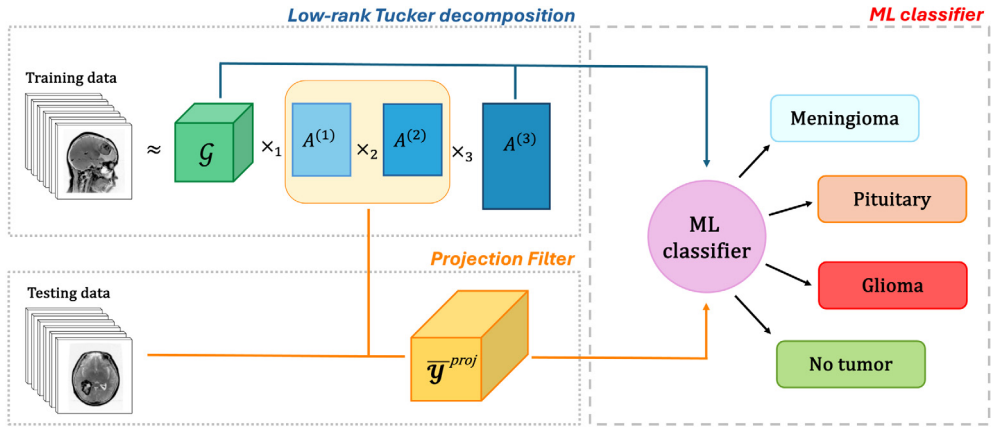


Fig. 7. The conceptual workflow illustrating a classification procedure based on the Tucker decomposition.

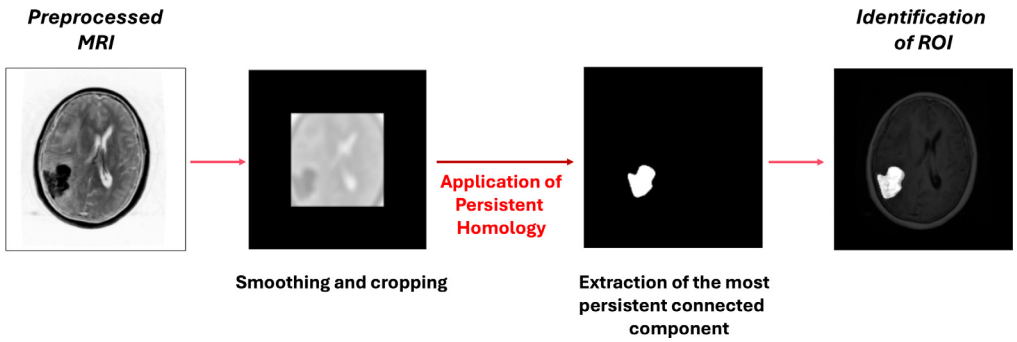


Fig. 8. The illustration shows how the tumor location in the imaging area of interest is determined by persistent homology.

smoothing factor of $N = 10$ for this process. This procedure is valuable for improving the application of PH. Smoothing reduces the variability of the grayscale values, thus filtering out pixels that could lead to excessively noisy connected components. We then cropped the MRI scans using a crop value of 70 to exclude the most external connected component, which for the brain image is represented by the skullcap.

2. **Computation of PH**. The lower star image filtration is used to compute the PH on images (as described in Section 2.2.3).
3. **Extraction of the most persistent connected component**. We used a diagonal automatic thresholding algorithm to extract information about the most persistent connected component from the PD. This method is based on the theoretical guarantee that it is possible to distinguish points corresponding to the most significant topological invariant from topological noise (i.e., points near the diagonal) in the PD, provided that the diagram has a certain width band without points [33]. Additionally, any distance within this band is theoretically acceptable as a threshold parameter, so the algorithm selects a threshold based on the largest distance between two consecutive lifetimes within this band. Thus, a significant area inside the MRI scan that may contain the tumor or indicate the presence of brain abnormalities is identified by the information received, which corresponds to a group of pixels representing the most persistent connected component discovered in the image.

The proposed framework offers two main outputs for each MRI scan, which should be jointly evaluated: a ROI in the MRI and the prediction of the tumor type with its relative likelihood, which is obtained by training the ML classifiers. ROI alerts the domain expert for additional research by identifying a possible cancerous region in the brain.

4. Results and discussion

This section presents and analyzes the experimental results obtained from the proposed framework. All the numerical experiments were carried out on the same workstation with an AMD Ryzen 7 CPU, 64 GB RAM, and without an external GPU.

The performances of the selected ML models were evaluated using accuracy, precision, recall and F1-score, defined as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{4}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

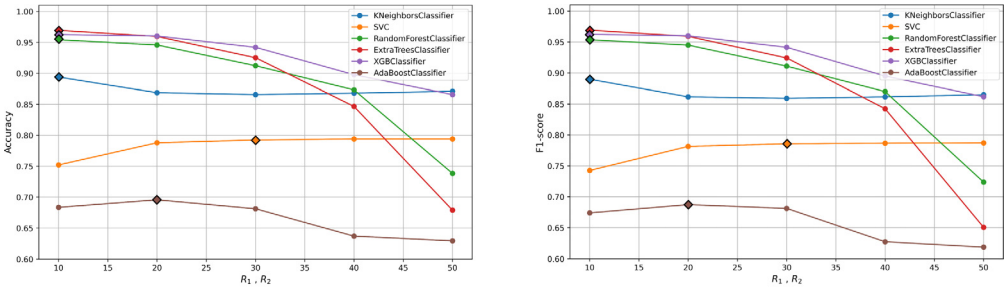


Fig. 9. Accuracy (a) and F1 (b) scores obtained by ML models depending on the dimension of mode 1 and mode 2 of core tensor. Mode 3 is fixed at 300. The selection of the optimal rank is highlighted with a diamond.

Table 2
Classification report obtained by Extra-Trees model on low-rank representation of testing set.

Class	Precision	Recall	F1-score	Support
Glioma	0.989	0.900	0.942	300
Meningioma	0.930	0.980	0.954	306
No tumor	0.997	1.000	0.998	405
Pituitary	0.967	0.997	0.981	300
Accuracy			**0.972**	1311
Weighted avg	0.972	0.971	**0.972**	1311

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

where TP, TN, FP, and FN correspond to the true positive, true negative, false positive, and false negative values, respectively. For all these metrics, higher values denotes a better detection and classification result; that is, predicted classes are more similar to ground truth ones. Confusion matrices are also provided.

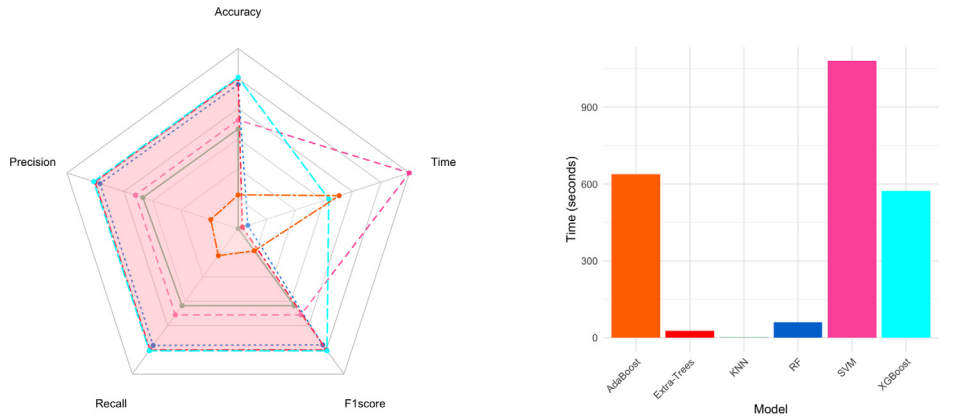
To demonstrate the benefits of starting with a low-rank representation of the data, we trained the ML models on both the original full-dimension input tensor ($250 \times 250 \times 5712$) and the projected core tensor obtained after applying TD ($R_1 \times R_2 \times R_3$). For each model, the value of R_3 was fixed at 300 after an extensive experimental session that took into account the trade-off between the accuracy of the data representation, the computational complexity of the models, and the required memory. Instead, we set the core dimension R_1 and R_2 to be equal and selected by maximizing the performance metrics as shown in Figs. 9(a) and 9(b).

Each ML method was trained five times to ensure a fair comparison. Figs. 10(a) and 10(b) (left panels) display radar charts of the average evaluation metrics for each model. The Extra-Trees and XGBoost models clearly outperform the others in terms of ML performance evaluation, regardless of whether the models are applied to the full-dimensional tensor or its reduced representation (Supplementary Tables S1 and S2).

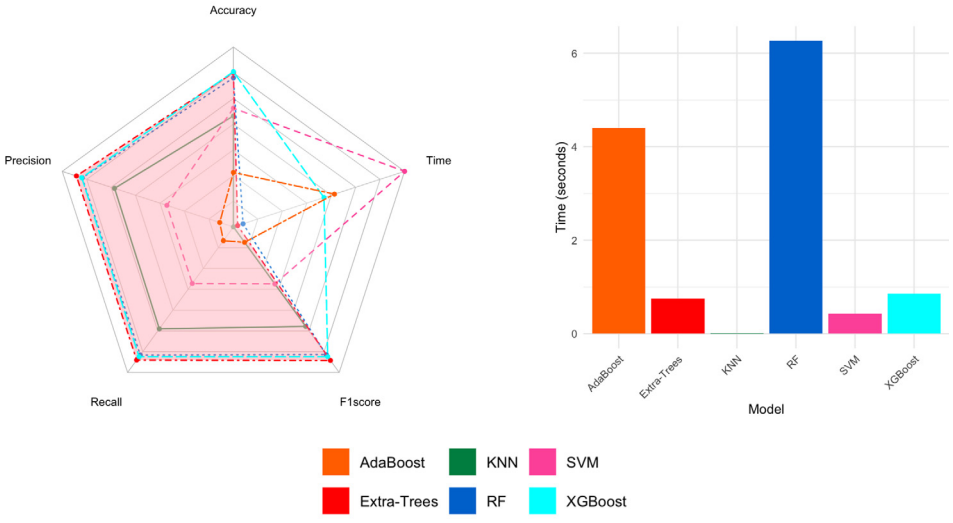
However, when applied to full-dimensional brain images, the XGBoost model requires significantly more training time than the Extra-Trees model (as it can be observed in Fig. 10(a), right panel). As expected, the use of TD reduces training times. In particular, the computational times of the models decrease significantly (Fig. 10(b) right panel) when using the compressed representation of the input data tensor. At the same time, the classification performance of some ML models is improved (Fig. 11). Specifically, in every metric taken into consideration, the Extra-Trees model outperforms XGBoost (Supplementary Table S2)). Extra-Trees was selected as the main model for the suggested framework because it produced the highest quantitative results on the low-rank representation of the data by optimizing the performance metrics and minimizing computational times. The core dimensions R_1 and R_2 are set equal to 10 by optimizing performance measures. The confusion matrix and classification report produced by the Extra-Trees model are shown in Fig. 12 and Table 2, respectively.

Through a comparative analysis conducted among previous studies that use the same dataset or subsets thereof, it results that our proposed ML-based methodology has demonstrated enhanced effectiveness in the classification of brain tumors, also encompassing deep learning architectures, as delineated in Table 3. The dimensionality reduction technique applied prior to the classifiers is highly innovative as it can significantly reduce the computational time and resource requirements while maintaining high accuracy. By employing this strategy, we are able to streamline the process without the need for extensive training data or long processing times associated with deep learning models. Our approach may offer advantages in scenarios where computational resources are limited or where rapid processing is essential, such as in clinical practice.

In contrast to conventional computational approaches, the proposed framework enhances model interpretability by combining ML-based methods with innovative topology-based techniques. Specifically, PH aims to facilitate the identification of ROI within MRI



(a)



(b)

Fig. 10. Radar chart displaying average performance metrics and computational times for all ML models trained on the original full-dimensional (a) and the low-rank representation (b) testing set, with the polygon area highlighted in red to indicate the model achieving superior metrics and lower computational times (left panel). Bar plot illustrating detailed computational times (right panel).

scans regardless of axial orientation or tumor location, thereby advancing understanding of complex medical images and supporting the clinical decision-making process. Fig. 13 shows examples of the outputs produced by our TDA-based method for each type of tumor observed in MRI scans acquired from various planes. The proposed approach successfully detects the tumor mass as well as the surrounding tumor microenvironment, emphasizing how crucial it is to identify and examine these areas to comprehend the biology of cancer, its causes, mutations, and the creation of effective therapies [38].

Furthermore, since the computation of persistent connected components relies on the intrinsic topology of the data, the size of the region of interest does not affect the computation of PH. In fact, our proposed approach is able to highlight both volumetrically substantial tumors (e.g., meningiomas in Fig. 13(a)), smaller tumors (e.g., gliomas in Fig. 13(b)), and tumors located within glands (e.g., pituitary in Fig. 13(c)). Moreover, the effectiveness of our approach is also evident when analyzing MRI scans without tumor mass. Fig. 14 illustrates that the most persistent connected component identified in no tumor images does not highlight a specific

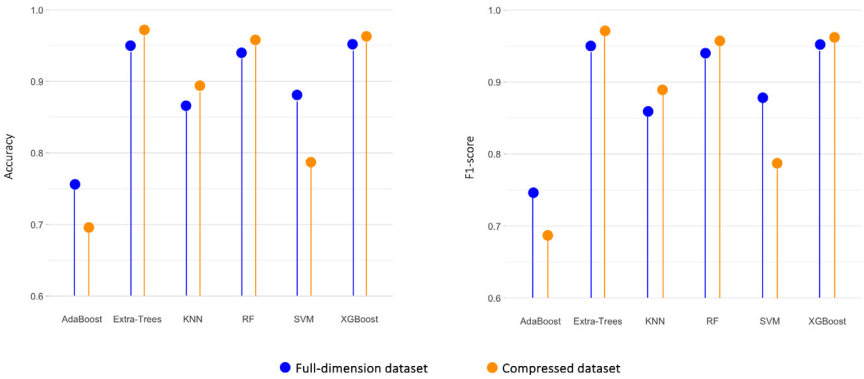


Fig. 11. Lollipop plot of accuracy (left panel) and F1 score (right panel) for models applied to the full-dimensional tensor and its reduced representation.

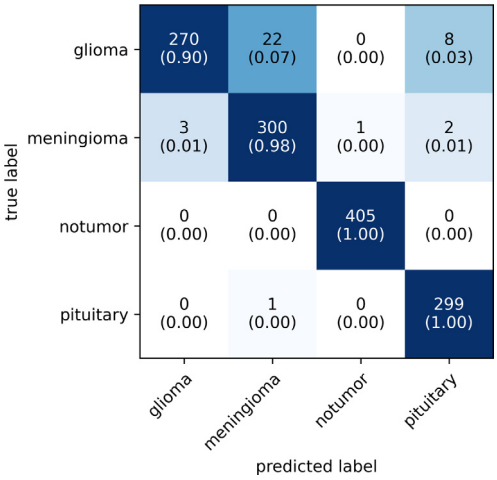


Fig. 12. Confusion matrix obtained by Extra-Trees model on low-rank representation of testing set.

Table 3
Comprehensive analysis and comparison of the obtained and previous studies' results. The asterisk '*' in the classes column denotes the use of a subset of the dataset images.

Authors	Year	Methods	Classes	Accuracy	F1-Score
Deepak et al. [5]	2020	CNN features + SVM	4	95.82	–
Filatov et al. [34]	2022	pretrained CNNs	4	89.55	–
Ekong et al. [9]	2022	Bayesian-CNN	4*	94.32	94.00
Almalki et al. [6]	2022	SURF + KAZE + SVM	4*	95.33	–
Chitnis et al. [35]	2022	LeaSE + DARTS architecture	4*	90.61	91.48
Ravinder et al. [36]	2023	CNN + GNN	4*	95.01	–
Gómez-Guzmán et al. [10]	2023	pretrained CNNs	4	97.12	–
Shilaskar et al. [37]	2023	HOG features + XGBoost	4	92.02	91.85
Proposed approach	**–**	**Tucker decomp. + Extra-Trees**	**4**	**97.28**	**97.16**

ROI in the MRI scans. The capability to identify and highlight tumors of various sizes, irrespective of their location or stage of progression, adds a great value for diagnostic purposes.

5. Conclusions

In this work, we present a novel ensemble approach in the field of medical imaging for brain tumor detection and classification. The primary goal of our methodology is to propose an automatic detection tool to assist domain experts in routine clinical practice. Our approach achieves this by integrating TD for dimensionality reduction and ML classifiers for brain tumor type prediction, while employing PH – a TDA technique – to identify ROIs in MRI scans. For the classification task, the results of our research study demonstrate a remarkable level of accuracy and F1-score, precisely 97.28% and 97.16%, achieved by training the Extra-Trees model. In addition, the innovative application of TDA, which exploits the geometric properties of MRI data, facilitates the

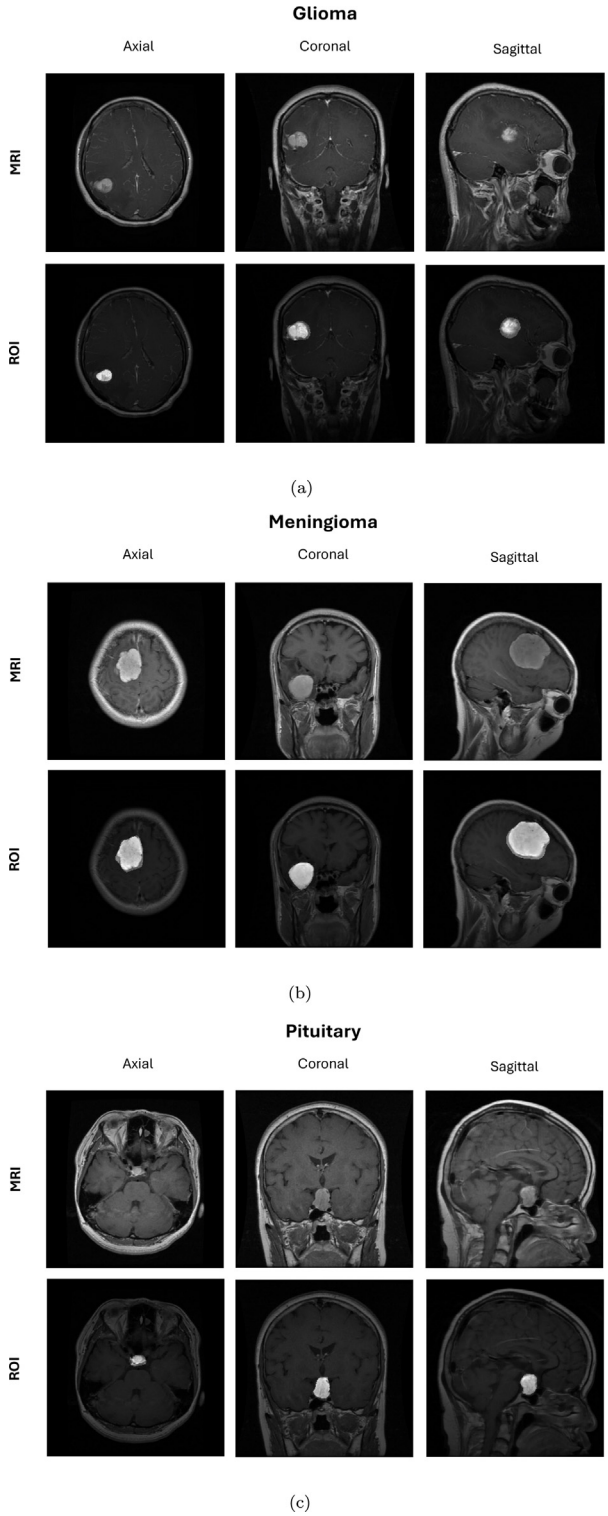


Fig. 13. Original MRI scans in the dataset (first rows) and the corresponding output ROIs by the proposed TDA-based methodology (second rows) for each type of the considered tumors: (a) glioma, (b) meningioma, and (c) pituitary.

identification of ROIs to discriminate the tumor mass and parts of the morphologically altered surrounding area. This comprehensive approach improves the interpretability of the classification process and could also potentially provide deeper insights into tumor

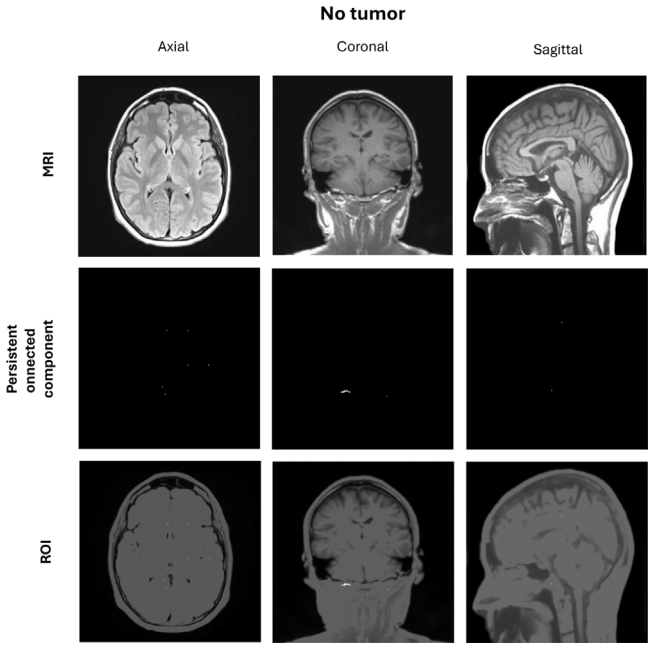


Fig. 14. No tumor MRI brain scans with its most persistent connected component and corresponding ROI.

characteristics. Given its current performance, our proposed approach emerges as a potentially promising automated approach to support clinicians in the detection and classification of brain tumors, particularly in suspected patients, and may also be applicable to other solid tumors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors S.G.D., G.G., and G.S. are members of the Gruppo Nazionale Calcolo Scientifico - Istituto Nazionale di Alta Matematica (GNCS-INdAM). G.G. and G.S. are supported by GNCS Project "Modelli con rango basso e algoritmi di ottimizzazione per l'analisi dati", Italy (CUP E53C23001670001). G.S. is funded by a PhD fellowship within the framework of the Italian "D.M. n. 352, April 9, 2022" - under the National Recovery and Resilience Plan, Msn. 4, Comp. 2, Investment 3.3 - PhD Project "Low-rank models for the analysis of Earth Observation data focusing on coastal and marine environments", co-supported by "Planetek Italia S.r.l". (CUP H91I22000410007). S.G.D. is funded by a PhD fellowship within the framework of the Italian "D.M. n. 117, March 2, 2023" - under the National Recovery and Resilience Plan, Msn. 4, Comp. 2, Investment 3.3 - PhD Project "Topological Data Analysis e tecniche di ottimizzazione per processi industriali", co-supported by "Pirelli Tyre S.p.A", Italy. (CUP H91I23000170007).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jcmds.2024.100103.

References

[1] Hussain S, et al. Modern diagnostic imaging technique applications and risk factors in the medical field: A review. BioMed Res Int 2022;2022:1–19. http://dx.doi.org/10.1155/2022/5164970.
[2] National Research Council. Mathematics and physics of emerging biomedical imaging. National Academies Press; 1996, http://dx.doi.org/10.17226/5066.
[3] Mabray MC, Barajas RF, Cha S. Modern brain tumor imaging. Brain Tumor Res Treat 2015;3(1):8. http://dx.doi.org/10.14791/btrt.2015.3.1.8.
[4] Lapointe S, Perry A, Butowski NA. Primary brain tumours in adults. Lancet 2018;392(10145):432–46.
[5] Deepak S, Ameer PM. Automated categorization of brain tumor from MRI using CNN features and SVM. J Ambient Intell Humaniz Comput 2020;12(8):8357–69. http://dx.doi.org/10.1007/s12652-020-02568-w.
[6] Almalki YE, et al. Robust Gaussian and nonlinear hybrid invariant clustered features aided approach for speeded brain tumor diagnosis. Life 2022;12(7):1084. http://dx.doi.org/10.3390/life12071084.

[7] Sekhar A, et al. Brain tumor classification using fine-tuned GoogLeNet features and machine learning algorithms: IoMT enabled CAD system. IEEE J Biomed Health Inf 2022;26(3):983–91. http://dx.doi.org/10.1109/jbhi.2021.3100758.
[8] Rasheed Z, et al. Brain tumor classification from MRI using image enhancement and convolutional neural network techniques. Brain Sci 2023;13(9):1320. http://dx.doi.org/10.3390/brainsci13091320.
[9] Ekong F, et al. Bayesian depth-wise convolutional neural network design for brain tumor MRI classification. Diagnostics 2022;12(7):1657. http://dx.doi.org/10.3390/diagnostics12071657.
[10] Gómez-Guzmán MA, et al. Classifying brain tumors on magnetic resonance imaging by using convolutional neural networks. Electronics 2023;12(4):955. http://dx.doi.org/10.3390/electronics12040955.
[11] Ferdous GJ, et al. LCDEiT: A linear complexity data-efficient image transformer for MRI brain tumor classification. IEEE Access 2023;11:20337–50. http://dx.doi.org/10.1109/access.2023.3244228.
[12] Tummala S, Kadry S, Bukhari SAC, Rauf HT. Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling. Curr Oncol 2022;29(10):7498–511. http://dx.doi.org/10.3390/curroncol29100590.
[13] Akinyelu AA, Zaccagna F, Grist JT, Castelli M, Rundo L. Brain tumor diagnosis using machine learning, convolutional neural networks, capsule neural networks and vision transformers, applied to MRI: A survey. J Imaging 2022;8(8):205. http://dx.doi.org/10.3390/jimaging8080205.
[14] Kolda TG, Bader BW. Tensor decompositions and applications. SIAM Rev 2009;51(3):455–500. http://dx.doi.org/10.1137/07070111x.
[15] Bhuvaji S, et al. Brain tumor classification (MRI): SARTAJ. Kaggle; 2020, http://dx.doi.org/10.34740/KAGGLE/DSV/1183165.
[16] Cheng J. Brain tumor dataset: figshare dataset. Figshare 2017. http://dx.doi.org/10.6084/m9.figshare.1512427.v5.
[17] Brain tumor classification (MRI): Br35h dataset. 2022, Kaggle, URL www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection?select=no.
[18] Weber M-A, Giesel FL, Stieltjes B. MRI for identification of progression in brain tumors: from morphology to function. Expert Rev Neurotherapeutics 2008;8(10):1507–25. http://dx.doi.org/10.1586/14737175.8.10.1507.
[19] De Lathauwer L, De Moor B, Vandewalle J. On the best rank-1 and rank-(R1 , R2 , . . ., RN) approximation of higher-order tensors. SIAM J Matrix Anal Appl 2000;21(4):1324–42. http://dx.doi.org/10.1137/s0895479898346995.
[20] Dasarathy BV. Nearest neighbor (NN) norms: NN pattern classification techniques. IEEE Comput Soc Tutor 1991.
[21] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273–97. http://dx.doi.org/10.1007/bf00994018.
[22] Breiman L. Random forests. Mach Learn 2001;45(1):5–32. http://dx.doi.org/10.1023/a:1010933404324.
[23] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn 2006;63(1):3–42. http://dx.doi.org/10.1007/s10994-006-6226-1.
[24] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM; 2016, http://dx.doi.org/10.1145/2939672.2939785.
[25] Hastie T, Rosset S, Zhu J, Zou H. Multi-class AdaBoost. Stat Interface 2009;2(3):349–60. http://dx.doi.org/10.4310/sii.2009.v2.n3.a8.
[26] Vandaele R, Nervo GA, Gevaert O. Topological image modification for object detection and topological image processing of skin lesions. Sci Rep 2020;10(1). http://dx.doi.org/10.1038/s41598-020-77933-y.
[27] Dey TK, Wang Y. Computational topology for data analysis. Cambridge University Press; 2022, http://dx.doi.org/10.1017/9781009099950.
[28] Singh Y, et al. Topological data analysis in medical imaging: Current state of the art. Insights Imaging 2023;14(1). http://dx.doi.org/10.1186/s13244-023-01413-w.
[29] Schenck H. Algebraic foundations for applied topology and data analysis. Springer International Publishing; 2022, http://dx.doi.org/10.1007/978-3-031-06664-1.
[30] Bleile B, Garin A, Heiss T, Maggs K, Robins V. The persistent homology of dual digital image constructions. 2021, http://dx.doi.org/10.48550/Arxiv.2102.11397, arXiv.
[31] Cohen-Steiner D, Edelsbrunner H, Harer J. Stability of persistence diagrams. Discrete Comput Geom 2006;37(1):103–20. http://dx.doi.org/10.1007/s00454-006-1276-5.
[32] Tahir B, et al. Feature enhancement framework for brain tumor segmentation and classification. Microsc Res Tech 2019;82(6):803–11. http://dx.doi.org/10.1002/jemt.23224.
[33] Chazal F, Guibas LJ, Oudot SY, Skraba P. Persistence-based clustering in Riemannian manifolds. In: Proceedings of the twenty-seventh annual symposium on computational geometry. ACM; 2011, http://dx.doi.org/10.1145/1998196.1998212.
[34] Filatov D, Ahmad Hassan Yar GN. Brain tumor diagnosis and classification via pre-trained convolutional neural networks. medRxiv 2022. http://dx.doi.org/10.1101/2022.07.18.22277779.
[35] Chitnis S, Hosseini R, Xie P. Brain tumor classification based on neural architecture search. Sci Rep 2022;12(1). http://dx.doi.org/10.1038/s41598-022-22172-6.
[36] Ravinder M, et al. Enhanced brain tumor classification using graph convolutional neural network architecture. Sci Rep 2023;13(1). http://dx.doi.org/10.1038/s41598-023-41407-8.
[37] Shilaskar S, et al. Machine learning based brain tumor detection and classification using HOG feature descriptor. In: 2023 international conference on sustainable computing and smart systems. IEEE; 2023, http://dx.doi.org/10.1109/icscss57650.2023.10169700.
[38] Azzarelli R, Simons BD, Philpott A. The developmental origin of brain tumours: A cellular and molecular framework. Development 2018;145(10). http://dx.doi.org/10.1242/dev.162693.